x00111602

IT TALLAGHT  Conor Griffin

CA2: Detect patients that have asd

report by: conor griffin x00111602

Contents

# Section 1

## Analyse the dataset

From looking at the dataset

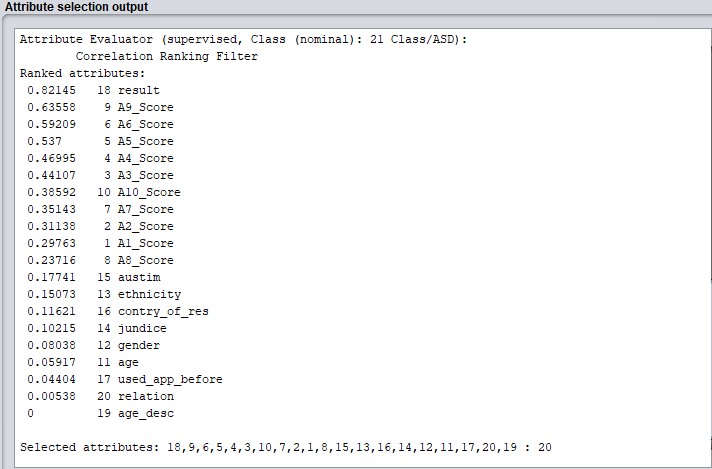
## Prepare a number of views (formats) of the dataset

## Attribute Selection (Feature selection)

For attribute selection I divided the process into two parts: Attribute evaluator, this is used to evaluate the attribute in the context of the class variable. The Search Method, the technique to try or navigate combinations of attributes in the dataset in order to get on a short list of chosen features.

The first attribute selection method I used to determine which attributes to choose is the correlation attribute evaluation technique. I used this method with a Ranker search method which evaluates each attribute and lists the results in rank order. I configured both the attribute evaluator and search method to work with each other in Weka. I ran the algorithm on the dataset and go the following results.

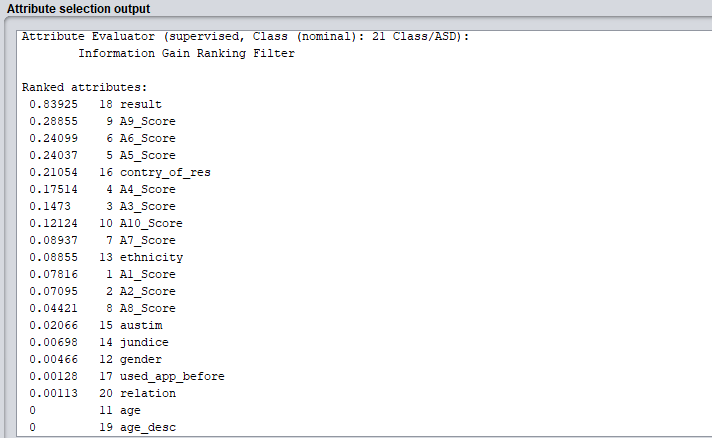




The technique to use here is to take a look at the correlation in attributes. Correlation is calculated for each variable predictor variable and the process to follow is to select only those attributes that have a moderate to high positive or negative correlation (close to -1 or 1) and drop those attributes with a low correlation (values close to 0). The use of the ranker search method also displays the attributes in a ranked order. Looking at the output you can see that the predictor attribute result has a really high correlation (0.82145) with the output class, also A9\_Score, A6\_score, A5\_Score, A4\_Score and A3\_Score have a high correlation with the class/ASD variable. If we set the cut-off point at 0.4 for relevant attributes then the remaining 14 attributes could possibly be removed, but this seems like a lot of data to remove so I would either set the cut-off lower or use other techniques to make a final decision on what attributes to remove.

The second method I used for attribute selection is the Information gain technique, which is used to calculate the info gain for each attribute for the output variable ASD. Entry values from 0 (no information) to 1 (maximum information). Those attributes that contribute more information will have a higher information gain value and can be selected whereas those that do not add much information will have a lower score and can be removed. In Weka I use the InfoGainAttributeEval in attribute evaluator and again use the Ranker search method. The following output displays the results.

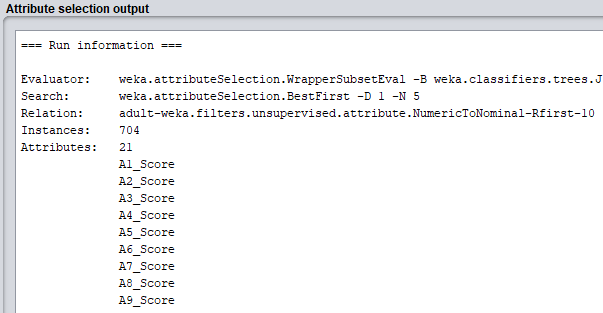


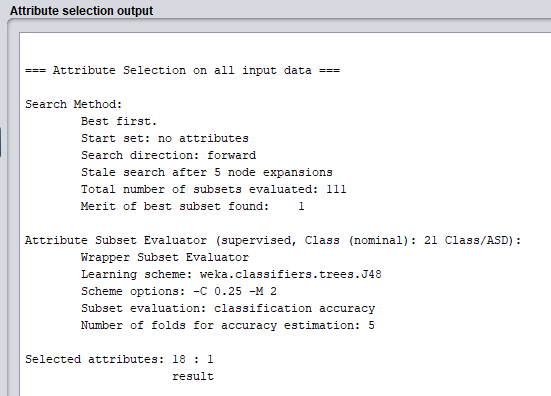


From the output you can see that the top ranked attribute is Result with a high information gain of 0.83925, most of the other attributes have a low info gain, but if we set the cut-off point at 0.2 then we could keep other attributes like A9, A6, A5\_Score and contry\_of\_res and the rest could be removed from the dataset.

The third selection method I used to determine what attributes to use in the modelling stage is the learner based feature selection. This is a powerful learning algorithm and is used to gather subsets of attributes, the subset with the best results and performance is taken as the selected subset. The feature selection method used is the WrapperSubsetEval technique and uses a BestFirst search method as it uses less compute time. I modified the configuration to instead of ZeroR, use J48 in trees, this gives a better preferred subset in the output.

The run information for this algorithm tells us that it ran the subset method with a best first search and used 704 instances over 21 attributes, it found 111 subsets and the merit of the best subset found was 1. The final selected attributes was attribute 18 result.





Looking back over the three techniques, you can see that a few of the attributes overlapped into each of the methods used. The results attribute would be the strongest and best attribute to keep in the dataset based on the three techniques used as it was number 1 in the output of each algorithm. Most of the other attributes didn’t perform well in each of the feature selection methods but this doesn’t necessarily mean you can get rid of all them, I would say it best to keep attributes that performed well in the first two techniques like A9\_Score, A6\_score, A5\_Score, A4\_Score and A3\_Score as well as contry\_of\_res attributes. I would add each of these in the modelling stage.

## Use kNearestNeighbor (IBK) Classifier on the dataset